Simple Linear Regression: Best Fit and Summary Formulas

BIOS 6611

CU Anschutz

Week 7

The Best Fit

The Regression Line

The line $Y = \beta_0 + \beta_1 X_1 + \epsilon$ is known as a **regression model**. The components include:

- β_0, the **intercept** of the line
- β_1, the **slope** of the line
- ϵ, the **error term** (i.e., the difference between the observed value and the (*unobserved*) *true* value of a quantity of interest, such as the population mean)

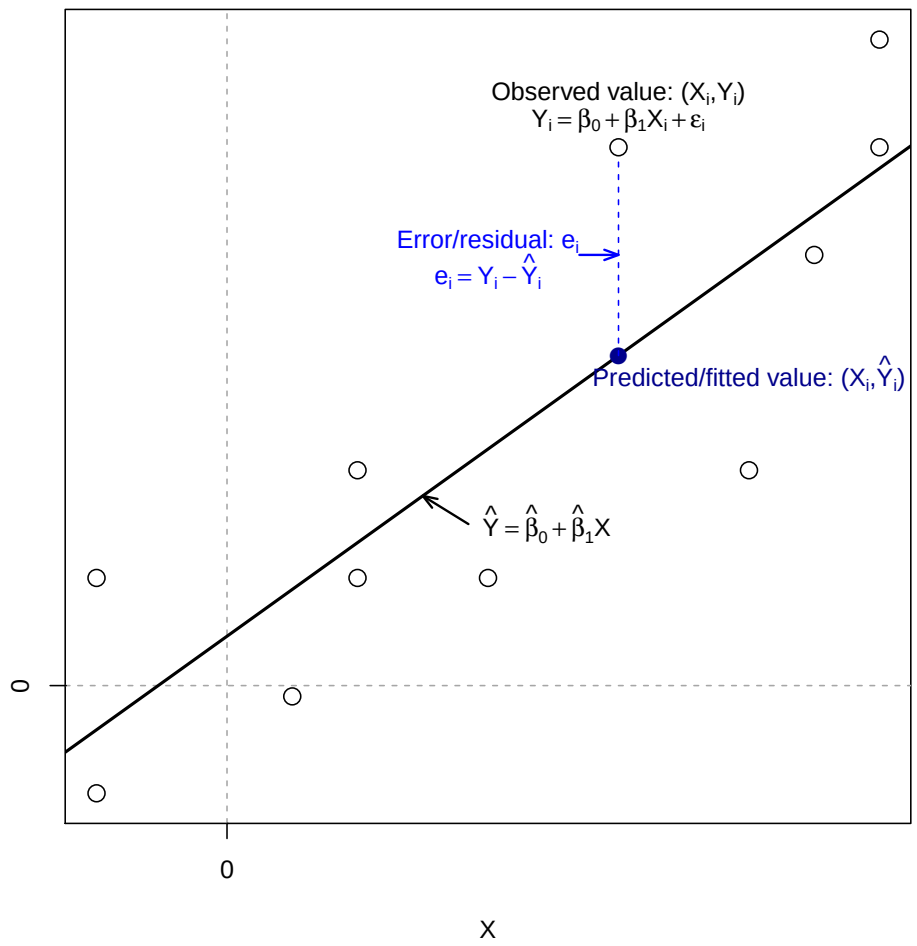
The predicted value of Y for a given value of X is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1,$$

where the hats represent the estimated values for our regression coefficients and the resulting FEV predicted value.

The Regression Line and Its Components

- Let (X_i, Y_i) be data for the ith individual for $i = 1, ..., n$.
- The difference between a fitted and observed value is called the **residual** (e_i), which is our sample estimate of the **error** (ϵ_i).
- We can choose the "best" line as the line that minimizes the residuals.



Approaches to Quantify the Total "Error"

Remember, the *error* (ϵ_i) is the unobserved true deviation of an observation from our quantity of interest. However, we can use the observable residuals (e_i) to quantify the total "error" and identify the "best" fit.

There are multiple approaches to quantify the total error:

1. $S = \sum_{i=1}^{n} e_i$ (i.e., the **sum of residuals**)
2. $S = \sum_{i=1}^{n} |e_i|$ (i.e., the **sum of the absolute value of residuals**)
3. $S = \sum_{i=1}^{n} e_i^2$ (i.e., the **sum of squares due to error**)

However, these are not all equally useful approaches:

1. An infinite number of approaches can minimize Approach 1.
2. Approach 2 is analytically difficult to work with.
3. Approach 3 is "easy" to use *and* has theoretical justification. This has become the standard method used to minimize the residuals.

Method of Least Squares/Least Squares Regression

$S = \sum_{i=1}^{n} e_i^2$ is called the **method of least squares** or **least squares regression** because it minimizes the sum of squares due to error (SS_{Error}):

$$SS_{Error} = SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right)^2$$

The SSE is also known as the **sums of squares error** or **residual sum of squares**.

Mathematical Approach to Least Squares

Mathematically stated, this approach identifies estimates for β_0 and β_1, $\hat{\beta}_0$ and $\hat{\beta}_1$, such that for any other possible estimators, $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$, it must be true that:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right)^2 < \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0^* + \hat{\beta}_1^* X_i) \right)^2$$

How can we arrive at these optimal estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$?

One approach is to treat our SS_{Error} as a *loss function* and minimize it over all choices for β_0 and β_1. To obtain the minimum (or maximum) of a function we find values such that the first (partial) derivatives are equal to 0. We will derive these in a separate slide set.

Summary Formulas

Formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$

All simple linear regression parameters can be estimates from 5 summary statistics:

1. n
2. $\sum X_i$ (can determine $\bar{X}$ by dividing by n)
3. $\sum Y_i$ (can determine $\bar{Y}$ by dividing by n)
4. $\sum(X_i^2)$
5. $\sum(X_i Y_i)$

With these 5 statistics we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

S_{XY} is the sums of squares of our cross-product between X and Y. S_{XX} is the sums of squares of X (which is connected to our sample variance if we divide it by $n-1$).

Relationships of $\hat{\beta}_1$ with Other Quantities

There are also some interesting connections for the sums of squares that make up $\hat{\beta}_1$:

$$\frac{S_{XY}}{S_{XX}} \times \frac{n-1}{n-1} = \frac{\frac{S_{XY}}{n-1}}{\frac{S_{XX}}{n-1}} = \frac{Cov(X,Y)}{Var(X)} = r_{X,Y}\frac{s_y}{s_x},$$

where s_y and s_x are the uncorrected sample standard deviations (i.e., they use n instead of $n-1$).

Variance Formulas

We also have formulas for the variance of our regression coefficients, which can help us to calculate confidence intervals or other summaries.

For these formulas, we first need to state the overall variance of Y:

$$\hat{\sigma}^2_Y = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$$

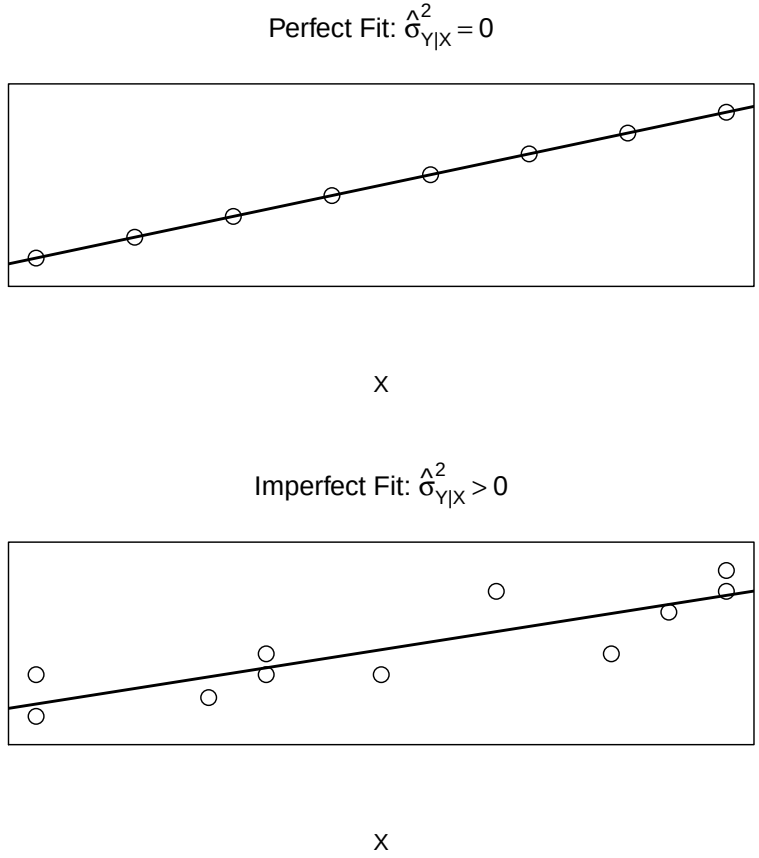
But for each value of X there is a subpopulation of values of Y. The variances of the subpopulations are $\sigma^2_{Y|X}$ are are estimated by:

$$\hat{\sigma}^2_{Y|X} = \frac{\sum_{i=1}^n (Y_i - \hat{Y})^2}{n-2} = \frac{\sum_{i=1}^n \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right)^2}{n-2} = \frac{SS_{Error}}{n-2} = MSE$$

This is also known as the **mean square error** or the **residual variance**.

$\hat{\sigma}^2_{Y|X}$ Behavior

- If $\hat{\sigma}^2_{Y|X}$ is 0, then the points will fall exactly on the regression line.
- If $\hat{\sigma}^2_{Y|X}$ is small, then the points will lie close to the regression line.
- If $\hat{\sigma}^2_{Y|X}$ is large, then the points will not fall close to the regression line. (Due to true variability in $Y|X$ and/or lack of fit.)
- The larger $\hat{\sigma}^2_{Y|X}$, the most scatter there will be in the data about the regression line.
- $\hat{\sigma}^2_{Y|X}$ is only an unbiased estimator for $\sigma^2_{Y|X}$ if the model is correct, otherwise $\hat{\sigma}^2_{Y|X} > \sigma^2_{Y|X}$.



Standard Errors for Our Intercept and Slope

If we assume a known mean square error (MSE):

$$SE(\hat{\beta}_0) = \sqrt{\sigma^2_{Y|X}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2_{Y|X}}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

If we estimate the MSE from our sample we can use the plug-in principle:

$$\hat{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2_{Y|X}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$$

$$\hat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Sampling Distributions for Our Intercept and Slope

Because the intercept and slope are statistics, they have their own sampling distributions.

If our assumptions (i.e., existence, linearity, independence, homoscedasticity, and normality of the errors) are met and we assume a known variance they will be normally distributed:

$$\hat{\beta}_0 \sim N\left(\beta_0, SE(\hat{\beta}_0)\right); \hat{\beta}_1 \sim N\left(\beta_1, SE(\hat{\beta}_1)\right)$$

Oftentimes we estimate σ^2 from a sample as s^2, and therefore the distribution will actually follow t_{n-2}. (Although as n gets larger it approximates the normal distribution.)

Specifically, we can make a connection to the standard normal (or t-distribution equivalent):

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1) \text{ or } \frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

Notation Summary

Notation Summary

Right Notation:

- Truth: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Expected: $E[Y_i] = \beta_0 + \beta_1 X_i$ since $E[\epsilon_i] = 0$
- Estimate: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Wrong Notation:

- $Y_i \neq \beta_0 + \beta_1 X_i$ (implies Y vs X is a perfect line)
- $E[Y_i] \neq \hat{\beta}_0 + \hat{\beta}_1 X_i$ since $E[\beta] = \beta$ and $E[\hat{\beta}] = \beta$
- Truth vs. estimate 1: $\hat{Y}_i \neq \beta_0 + \beta_1 X_i$
- Truth vs. estimate 2: $Y_i \neq \hat{\beta}_0 + \hat{\beta}_1 X_i$
- Truth vs. estimate 3: $\epsilon_i \neq e_i = Y_i - \hat{Y}_i$